Week 2 submission on Simple Linear Regression:

Submitted by: Tabassum Raizada

**About the dataset:**

?Boston

The Boston data frame has 506 rows and 14 columns. This data frame contains the following columns:

|  |  |
| --- | --- |
| crim | per capita crime rate by town. |
| zn | proportion of residential land zoned for lots over 25,000 sq.ft. |
| indus | proportion of non-retail business acres per town. |
| chas | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). |
| nox | nitrogen oxides concentration (parts per 10 million). |
| rm | average number of rooms per dwelling. |
| age | proportion of owner-occupied units built prior to 1940. |
| dis | weighted mean of distances to five Boston employment centres. |
| rad | index of accessibility to radial highways. |
| tax | full-value property-tax rate per \$10,000. |
| ptratio | pupil-teacher ratio by town. |
| black | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. |
| lstat | lower status of the population (percent). |
| medv | median value of owner-occupied homes in \$1000s |

> library(MASS)

> rm(list=ls(all=TRUE))

> # see the

> str(Boston)

'data.frame': 506 obs. of 14 variables:

$ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...

$ zn : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...

$ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...

$ chas : int 0 0 0 0 0 0 0 0 0 0 ...

$ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...

$ rm : num 6.58 6.42 7.18 7 7.15 ...

$ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...

$ dis : num 4.09 4.97 4.97 6.06 6.06 ...

$ rad : int 1 2 2 3 3 3 5 5 5 5 ...

$ tax : num 296 242 242 222 222 222 311 311 311 311 ...

$ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...

$ black : num 397 397 393 395 397 ...

$ lstat : num 4.98 9.14 4.03 2.94 5.33 ...

$ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

> summary(Boston)

crim zn indus chas nox

Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000 Min. :0.3850

1st Qu.: 0.08204 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000 1st Qu.:0.4490

Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000 Median :0.5380

Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917 Mean :0.5547

3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000 3rd Qu.:0.6240

Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000 Max. :0.8710

rm age dis rad tax

Min. :3.561 Min. : 2.90 Min. : 1.130 Min. : 1.000 Min. :187.0

1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100 1st Qu.: 4.000 1st Qu.:279.0

Median :6.208 Median : 77.50 Median : 3.207 Median : 5.000 Median :330.0

Mean :6.285 Mean : 68.57 Mean : 3.795 Mean : 9.549 Mean :408.2

3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188 3rd Qu.:24.000 3rd Qu.:666.0

Max. :8.780 Max. :100.00 Max. :12.127 Max. :24.000 Max. :711.0

ptratio black lstat medv

Min. :12.60 Min. : 0.32 Min. : 1.73 Min. : 5.00

1st Qu.:17.40 1st Qu.:375.38 1st Qu.: 6.95 1st Qu.:17.02

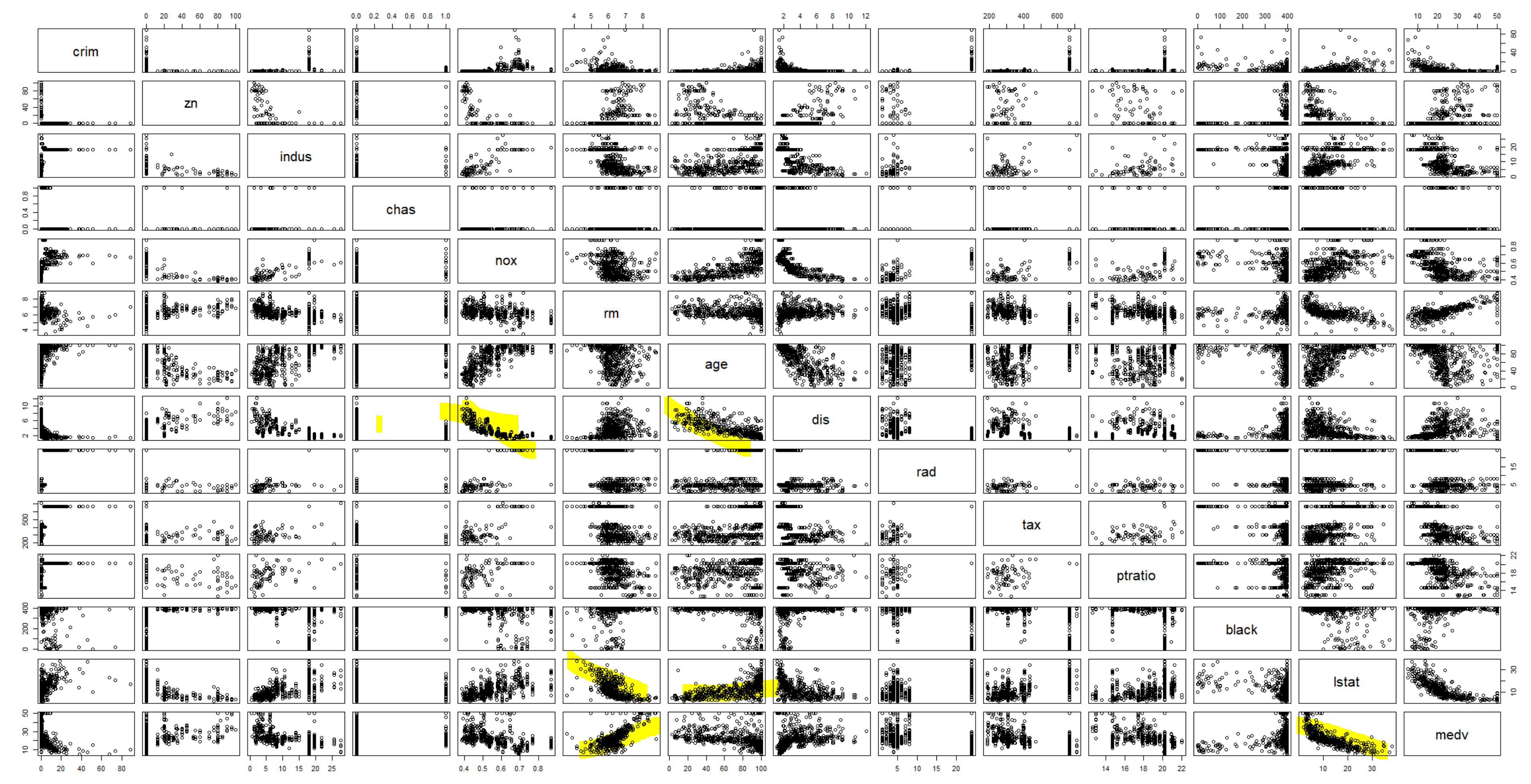
Median :19.05 Median :391.44 Median :11.36 Median :21.20

Mean :18.46 Mean :356.67 Mean :12.65 Mean :22.53

3rd Qu.:20.20 3rd Qu.:396.23 3rd Qu.:16.95 3rd Qu.:25.00

Max. :22.00 Max. :396.90 Max. :37.97 Max. :50.00

Some of the relationships that looked like have a linear relationship are highlighted below:



Relationship between the following variables can be looked into further:

NOX & DIS

AGE & DIS

AGE & LSTAT

AGE & MEDV

RM & LSTAT

RM & MEDV

Finding the correlation between the variables identified above:

> cor(Boston$age, Boston$lstat)

[1] 0.6023385

> cor(Boston$age, Boston$dis)

[1] -0.7478805

> cor(Boston$age, Boston$medv)

[1] -0.3769546

> cor(Boston$nox, Boston$dis)

[1] -0.7692301

> cor(Boston$rm, Boston$lstat)

[1] -0.6138083

> cor(Boston$rm, Boston$medv)

[1] 0.6953599

Some example of poor correlations:

> cor(Boston$crim, Boston$medv)

[1] -0.3883046

> cor(Boston$dis, Boston$tax)

[1] -0.5344316

Distance and Tax seems to have surprisingly high correlation ship than expected

### We will also compare results using the chart.Correlation() function in the package PerformanceAnalytics to draw scatter plots

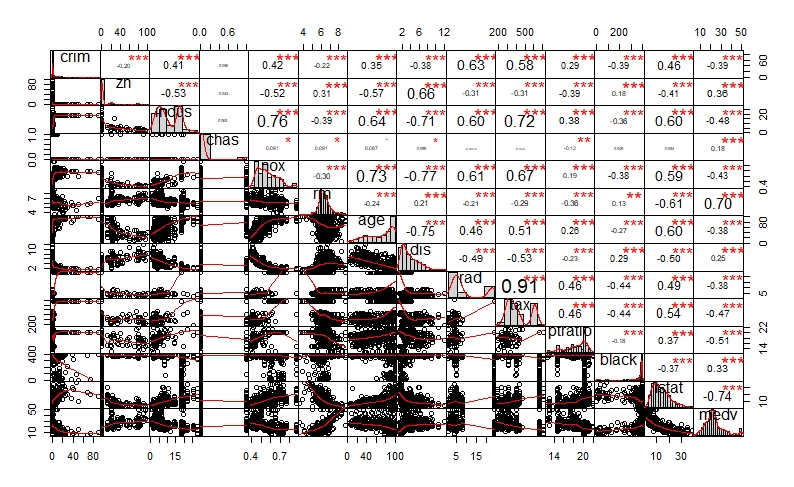
# Install PerformanceAnalytics:

install.packages("PerformanceAnalytics")

# Use chart.Correlation():

library("PerformanceAnalytics")

chart.Correlation(Boston, histogram=TRUE, pch=19)



The distribution of each variable is shown on the diagonal.

On the bottom of the diagonal : the bivariate scatter plots with a fitted line are displayed

On the top of the diagonal : the value of the correlation plus the significance level as stars

Each significance level is associated to a symbol : p-values(0, 0.001, 0.01, 0.05, 0.1, 1) <=> symbols(“\*\*\*”, “\*\*”, “\*”, “.”, " “)

**Determine correlation between variables:**

Correlation matrix with significance levels (p-value)

The function rcorr() [in Hmisc package] can be used to compute the significance levels for pearson and spearman correlations. It returns both the correlation coefficients and the p-value of the correlation for all possible pairs of columns in the data table

install.packages("Hmisc")

library("Hmisc")

res2 <- rcorr(as.matrix(Boston))

res2

The output of the function rcorr() is a list containing the following elements :

- r : the correlation matrix

crim zn indus chas nox rm age dis rad tax ptratio black lstat medv

crim 1.00 -0.20 0.41 -0.06 0.42 -0.22 0.35 -0.38 0.63 0.58 0.29 -0.39 0.46 -0.39

zn -0.20 1.00 -0.53 -0.04 -0.52 0.31 -0.57 0.66 -0.31 -0.31 -0.39 0.18 -0.41 0.36

indus 0.41 -0.53 1.00 0.06 0.76 -0.39 0.64 -0.71 0.60 0.72 0.38 -0.36 0.60 -0.48

chas -0.06 -0.04 0.06 1.00 0.09 0.09 0.09 -0.10 -0.01 -0.04 -0.12 0.05 -0.05 0.18

nox 0.42 -0.52 0.76 0.09 1.00 -0.30 0.73 -0.77 0.61 0.67 0.19 -0.38 0.59 -0.43

rm -0.22 0.31 -0.39 0.09 -0.30 1.00 -0.24 0.21 -0.21 -0.29 -0.36 0.13 -0.61 0.70

age 0.35 -0.57 0.64 0.09 0.73 -0.24 1.00 -0.75 0.46 0.51 0.26 -0.27 0.60 -0.38

dis -0.38 0.66 -0.71 -0.10 -0.77 0.21 -0.75 1.00 -0.49 -0.53 -0.23 0.29 -0.50 0.25

rad 0.63 -0.31 0.60 -0.01 0.61 -0.21 0.46 -0.49 1.00 0.91 0.46 -0.44 0.49 -0.38

tax 0.58 -0.31 0.72 -0.04 0.67 -0.29 0.51 -0.53 0.91 1.00 0.46 -0.44 0.54 -0.47

ptratio 0.29 -0.39 0.38 -0.12 0.19 -0.36 0.26 -0.23 0.46 0.46 1.00 -0.18 0.37 -0.51

black -0.39 0.18 -0.36 0.05 -0.38 0.13 -0.27 0.29 -0.44 -0.44 -0.18 1.00 -0.37 0.33

lstat 0.46 -0.41 0.60 -0.05 0.59 -0.61 0.60 -0.50 0.49 0.54 0.37 -0.37 1.00 -0.74

medv -0.39 0.36 -0.48 0.18 -0.43 0.70 -0.38 0.25 -0.38 -0.47 -0.51 0.33 -0.74 1.00

- n : the matrix of the number of observations used in analyzing each pair of variables

n= 506

- P : the p-values corresponding to the significance levels of correlations.

crim zn indus chas nox rm age dis rad tax ptratio black lstat medv

crim 0.0000 0.0000 0.2094 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

zn 0.0000 0.0000 0.3378 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

indus 0.0000 0.0000 0.1575 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

chas 0.2094 0.3378 0.1575 0.0403 0.0402 0.0518 0.0257 0.8687 0.4244 0.0062 0.2733 0.2259 0.0000

nox 0.0000 0.0000 0.0000 0.0403 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

rm 0.0000 0.0000 0.0000 0.0402 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0039 0.0000 0.0000

age 0.0000 0.0000 0.0000 0.0518 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

dis 0.0000 0.0000 0.0000 0.0257 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

rad 0.0000 0.0000 0.0000 0.8687 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

tax 0.0000 0.0000 0.0000 0.4244 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

ptratio 0.0000 0.0000 0.0000 0.0062 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

black 0.0000 0.0000 0.0000 0.2733 0.0000 0.0039 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

lstat 0.0000 0.0000 0.0000 0.2259 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

medv 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

We will use a simple function for formatting a correlation matrix into a table with 4 columns containing :

Column 1 : row names (variable 1 for the correlation test)

Column 2 : column names (variable 2 for the correlation test)

Column 3 : the correlation coefficients

Column 4 : the p-values of the correlations

flattenCorrMatrix <- function(cormat, pmat) {

ut <- upper.tri(cormat)

data.frame(

row = rownames(cormat)[row(cormat)[ut]],

column = rownames(cormat)[col(cormat)[ut]],

cor =(cormat)[ut],

p = pmat[ut]

)

}

res3<-rcorr(as.matrix(Boston))

unsorted <- flattenCorrMatrix(res3$r, res3$P)

> unsorted[order(unsorted$cor,decreasing=TRUE),]

row column cor p

45 rad tax 0.910228189 0.000000e+00

9 indus nox 0.763651447 0.000000e+00

20 nox age 0.731470104 0.000000e+00

39 indus tax 0.720760180 0.000000e+00

84 rm medv 0.695359947 0.000000e+00

41 nox tax 0.668023200 0.000000e+00

23 zn dis 0.664408223 0.000000e+00

18 indus age 0.644778511 0.000000e+00

29 crim rad 0.625505145 0.000000e+00

33 nox rad 0.611440563 0.000000e+00

69 indus lstat 0.603799716 0.000000e+00

73 age lstat 0.602338529 0.000000e+00

31 indus rad 0.595129275 0.000000e+00

71 nox lstat 0.590878921 0.000000e+00

37 crim tax 0.582764312 0.000000e+00

76 tax lstat 0.543993412 0.000000e+00

43 age tax 0.506455594 0.000000e+00

75 rad lstat 0.488676335 0.000000e+00

54 rad ptratio 0.464741179 0.000000e+00

55 tax ptratio 0.460853035 0.000000e+00

35 age rad 0.456022452 0.000000e+00

67 crim lstat 0.455621479 0.000000e+00

7 crim nox 0.420971711 0.000000e+00

2 crim indus 0.406583411 0.000000e+00

48 indus ptratio 0.383247556 0.000000e+00

77 ptratio lstat 0.374044317 0.000000e+00

80 zn medv 0.360445342 0.000000e+00

16 crim age 0.352734251 4.440892e-16

90 black medv 0.333460820 1.332268e-14

12 zn rm 0.311990587 6.936673e-13

63 dis black 0.291511673 2.278644e-11

46 crim ptratio 0.289945579 2.942935e-11

52 age ptratio 0.261515012 2.338885e-09

86 dis medv 0.249928734 1.206612e-08

27 rm dis 0.205246213 3.237746e-06

50 nox ptratio 0.188932677 1.885692e-05

57 zn black 0.175520317 7.207719e-05

82 chas medv 0.175260177 7.390623e-05

61 rm black 0.128068635 3.906695e-03

14 chas rm 0.091251225 4.018410e-02

10 chas nox 0.091202807 4.029050e-02

19 chas age 0.086517774 5.177446e-02

6 indus chas 0.062938027 1.574628e-01

59 chas black 0.048788485 2.733379e-01

32 chas rad -0.007368241 8.686789e-01

40 chas tax -0.035586518 4.244225e-01

5 zn chas -0.042696719 3.378103e-01

70 chas lstat -0.053929298 2.258990e-01

4 crim chas -0.055891582 2.094345e-01

25 chas dis -0.099175780 2.568848e-02

49 chas ptratio -0.121515174 6.203916e-03

66 ptratio black -0.177383302 6.017320e-05

1 crim zn -0.200469220 5.506472e-06

34 rm rad -0.209846668 1.918446e-06

11 crim rm -0.219246703 6.346703e-07

53 dis ptratio -0.232470542 1.229920e-07

21 rm age -0.240264931 4.459649e-08

62 age black -0.273533977 3.911800e-10

42 rm tax -0.292047833 2.086820e-11

15 nox rm -0.302188188 3.818723e-12

30 zn rad -0.311947826 6.987744e-13

38 zn tax -0.314563325 4.385381e-13

51 rm ptratio -0.355501495 0.000000e+00

58 indus black -0.356976535 0.000000e+00

78 black lstat -0.366086902 0.000000e+00

85 age medv -0.376954565 0.000000e+00

22 crim dis -0.379670087 0.000000e+00

60 nox black -0.380050638 0.000000e+00

87 rad medv -0.381626231 0.000000e+00

56 crim black -0.385063942 0.000000e+00

79 crim medv -0.388304609 0.000000e+00

13 indus rm -0.391675853 0.000000e+00

47 zn ptratio -0.391678548 0.000000e+00

68 zn lstat -0.412994575 0.000000e+00

83 nox medv -0.427320772 0.000000e+00

65 tax black -0.441808007 0.000000e+00

64 rad black -0.444412816 0.000000e+00

88 tax medv -0.468535934 0.000000e+00

81 indus medv -0.483725160 0.000000e+00

36 dis rad -0.494587930 0.000000e+00

74 dis lstat -0.496995831 0.000000e+00

89 ptratio medv -0.507786686 0.000000e+00

8 zn nox -0.516603708 0.000000e+00

3 zn indus -0.533828186 0.000000e+00

44 dis tax -0.534431584 0.000000e+00

17 zn age -0.569537342 0.000000e+00

72 rm lstat -0.613808272 0.000000e+00

24 indus dis -0.708026989 0.000000e+00

91 lstat medv -0.737662726 0.000000e+00

28 age dis -0.747880541 0.000000e+00

26 nox dis -0.769230113 0.000000e+00

The function corrplot() creates a graphical display of a correlation matrix, highlighting the most correlated variables in a data table. In this plot, correlation coefficients are colored according to the value. Correlation matrix can be also reordered according to the degree of association between variables.

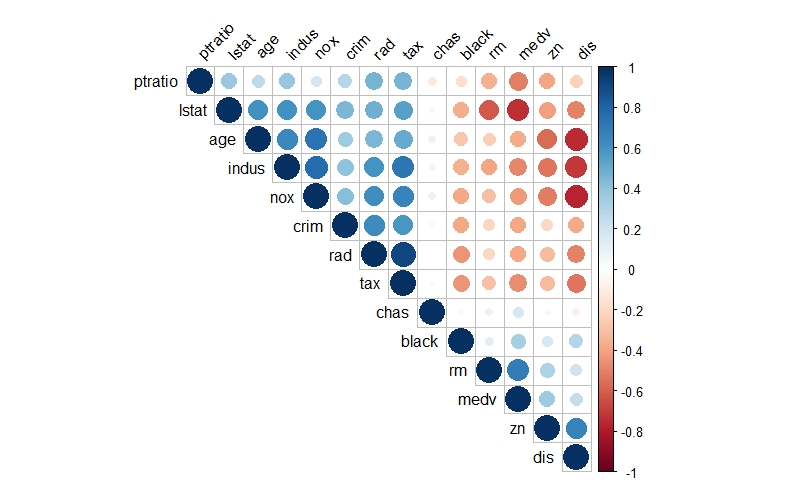
install.packages("corrplot")

#We use corrplot() to create a correlogram. The function corrplot() takes the correlation matrix as the first argument. The second argument (type=“upper”) is used to display only the upper triangular of the correlation matrix.

library(corrplot)

corrplot(res, type = "upper", order = "hclust",

tl.col = "black", tl.srt = 45)



**Building a Linear Regression Model:**

We will look further into the relationship between rm (average number of rooms per dwelling) & medv (Median value of owner-occupied homes in $1000's)

**Rm = β0 +β1 \* Medv + €**

**Statement of Null and the alternative hypothesis:**

**H0 (NULL Hypothesis) :** coefficient **β1** is equal to zero (no effect). Predictor Medv has no effect on the response variable Rm.

**β1 = 0**

**H1 (Alternate Hypothesis) : β1** is not equal to zero. changes in the predictor's value are related to changes in the response variable.

**β1 ~~≠~~ 0**

**Linear regression commands in R and the corresponding results**

attach (Boston )

lm.fit =lm(rm ~ medv)

> lm.fit

Call:

lm(formula = rm ~ medv)

Coefficients:

(Intercept) medv

5.08764 0.05312

For more detailed information, we use summary(lm.fit). This gives us pvalues and standard errors for the coefficients, as well as the R2 statistic and F-statistic for the model.

> summary(lm.fit)

Call:

lm(formula = rm ~ medv)

Residuals:

Min 1Q Median 3Q Max

-2.98750 -0.24448 0.01893 0.27379 2.52898

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.087639 0.059510 85.49 <2e-16 \*\*\*

medv 0.053122 0.002446 21.72 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5054 on 504 degrees of freedom

Multiple R-squared: 0.4835, Adjusted R-squared: 0.4825

F-statistic: 471.8 on 1 and 504 DF, p-value: < 2.2e-16

Model coefficients and significance test: The hypothesis test of the model coefficients and their significance tests provide information regarding to which coefficients are significantly different than 0 (asterisk signs) and these indicate that coefficients are contributing to the model.

In order to obtain a confidence interval for the coefficient estimates, we will use the confint() command.

> confint(lm.fit,level=0.95)

2.5 % 97.5 %

(Intercept) 4.97072001 5.20455733

medv 0.04831762 0.05792709

level = level of confidence required

**Accept or reject the null hypothesis**

The p-value for medv (<2e-16) is much less than the confidence level of 0.05, which indicates that it is statistically significant. Hence we reject the NULL hypothesis.  In other words, medv predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the medv value are related to changes in the response variable (rm).

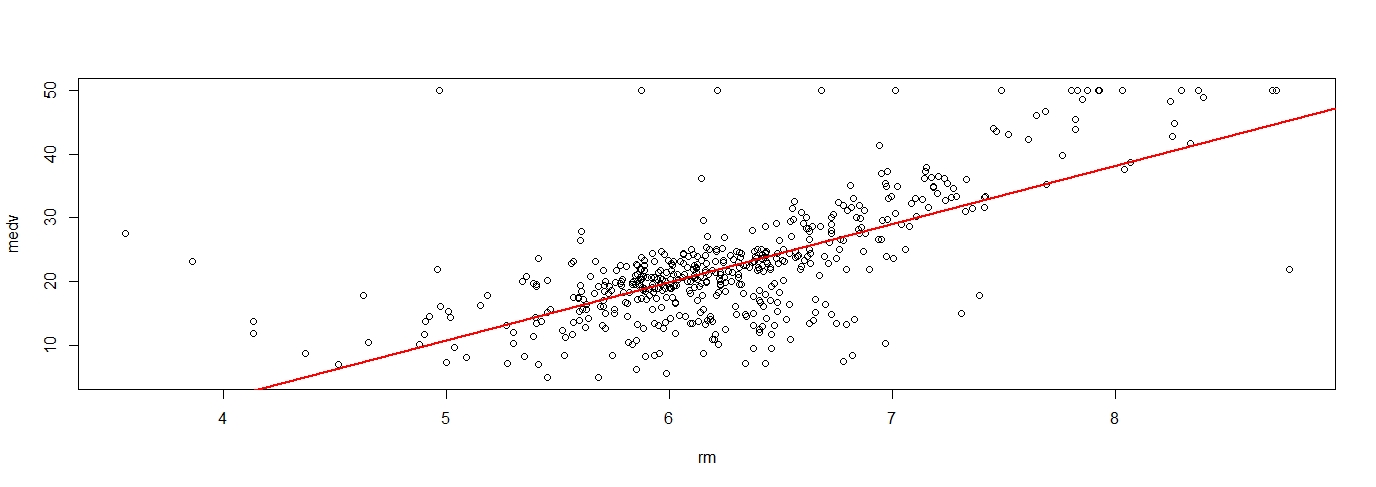
**Evidence of a linear relationship between the predictor and response variables**

We will plot rm & medv along the least square line using plot() and abline() functions.

The data does show outliers (medv values >5,000,000)

plot(rm,medv)

abline(lm.fit,lwd=2,col="red")

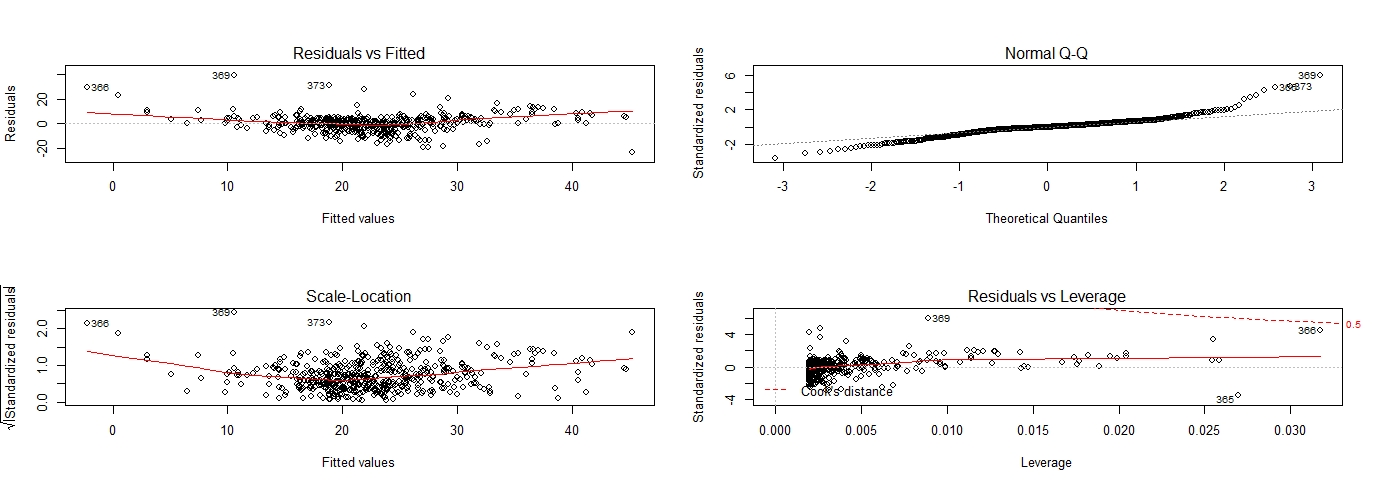


**Verify the model assumptions**

We now examine some diagnostic plots to verify model assumptions of linearity, normality, variance. Four diagnostic plots are automatically produced by applying the plot() function directly to the output from lm(). This command will produce one plot at a time, and hitting Enter generates the next plot. We use the par() function, to view all four plots together on a 2 × 2 grid of panels.

> par(mfrow =c(2,2))

> plot(lm.fit)



The first plot (residuals vs. fitted values) is a simple scatterplot between residuals and predicted values.  It should look more or less random. However a strong pattern in the residuals indicates non-linearity of data. Hence the model could benefit from non-linear transformations

#### The second plot (normal Q-Q) is a [normal probability plot](http://www.theanalysisfactor.com/anatomy-of-a-normal-probability-plot/).  It will give a straight line if the errors are distributed normally. We see a QQ plot where the residuals deviate from the diagonal line in both the upper and lower tail. We can observe a heavier tail in our case. this case we see that the tails are observed to be ‘heavier’ (have larger values) than what we would expect under the standard modeling assumptions. This is indicated by the points forming a “steeper” line than the diagonal.

Our assumption is that the errors are approximately normally distributed, which is what the Q-Q plot allows you to see.

The third plot (Scale-Location also called spread-location plot), like the first, should look random i.e we should see no patterns.  In the above case it is more U-shaped. This plot shows if residuals are spread equally along the ranges of predictors. This helps check the assumption of equal variance (homoscedasticity). It is ideal to see a horizontal line with equally (randomly) spread points. Even though the line is not straight the points are more or less evenly spread.

The last plot (Cook’s distance) tells us which points have the greatest influence on the regression (leverage points). 366 has a high leverage statistic along with having a high residual. Hence it is a outlier and high leverage observation.

**Summary of findings:**

1. The linear model (rm, mdev) can be improved by removing the outliers with high leverage
2. Not all the assumptions of the linear regression model are met:
   1. Error is not normally distributed
   2. The residual show that the linear model does not represent the data well. It reveals unexplained patterns in the data by the fitted model. Hence it is worth exploring the model using quadratic functions.
3. It is worth building the model using multiple variables to see the influence of other variables could have on the predicted value.